A nanoinformatics decision support tool for the virtual screening of gold nanoparticle cellular association using protein corona fingerprints

Afantitis, Antreas; Melagraki, Georgia; Tsoumanis, Andreas; Valsami-Jones, Eugenia; Lynch, Iseult

DOI:
10.1080/17435390.2018.1504998

Document Version
Peer reviewed version

Citation for published version (Harvard):
Afantitis, A, Melagraki, G, Tsoumanis, A, Valsami-Jones, E & Lynch, I 2018, 'A nanoinformatics decision support tool for the virtual screening of gold nanoparticle cellular association using protein corona fingerprints', *Nanotoxicology*, vol. 12, no. 10, pp. 1148-1165. https://doi.org/10.1080/17435390.2018.1504998

A NANOINFORMATICS DECISION SUPPORT TOOL FOR THE VIRTUAL SCREENING OF GOLD NANOPARTICLE CELLULAR ASSOCIATION USING PROTEIN CORONA FINGERPRINTS

Antreas Afantitis,[1] Georgia Melagraki,[1*] Andreas Tsoumanis[1], Eva Valsami-Jones,[2] Iseult Lynch[2]*

[1] Nanoinformatics Department, Novamechanics Ltd, Nicosia, Cyprus

[2] School of Geography, Earth and Environmental Sciences, University of Birmingham, B15 2TT Birmingham, United Kingdom.

Corresponding authors: melagraki@novamechanics.com; i.lynch@bham.ac.uk

Abstract

The increasing use of nanoparticles (NPs) in a wide range of consumer and industrial applications has necessitated significant effort to address the challenge of characterizing and quantifying the underlying nanostructure – biological response relationships to ensure that these novel materials can be exploited responsibly and safely. Such efforts demand reliable experimental data not only in terms of the biological dose-response but also regarding the physicochemical properties of the NPs and their interaction with the biological environment. The latter has not been extensively studied, as a large surface to bind biological macromolecules is a unique feature of NPs that is not relevant for chemicals or pharmaceuticals, and thus only limited data have been reported in the literature quantifying the protein corona formed when NPs interact with a biological medium and linking this with NP cellular association/uptake. In this work we report the development of a predictive model for the assessment of the biological response (cellular association, which can include both internalized NPs and those attached to the cell surface) of surface-modified gold NPs, based on their physicochemical properties and protein corona fingerprints, utilizing a dataset of 105 unique NPs. Cellular association was chosen as the end-point for the original experimental study due to its relevance to inflammatory responses, biodistribution, and toxicity *in vivo.* The validated predictive model is freely available online through the Enalos Cloud Platform (http://enalos.insilicotox.com/NanoProteinCorona/) for use as part of a regulatory or NP safe-by-design decision support system. This online tool will allow the virtual screening of NPs, based on a list of significant NP descriptors, identifying those NPs that would warrant further toxicity testing on the basis of predicted NP cellular association.

Keywords

Nanoparticles, Nanoinformatics, protein corona, cell association, Enalos Cloud platform, web service, toxicity, hazard characterization, risk assessment

1. Introduction

Nanoparticles (NPs) unique properties are increasingly gaining attention in a wide range of applications spanning from electronics, to drug delivery, defence applications and many more. A critical mass of scientific effort is currently directed towards NPs research and technology development targeting new improved structures for all possible uses. Recent progress in the field resulted in the emergence of several NPs as valuable alternatives to 'traditional' bulk materials and simultaneously raised concerns regarding the short and long term effects of such novel materials on human health and the environment.(Haase, Tentschert and Luch, 2012; D. Fourches, 2014; Tropsha, Mills and Hickey, 2017)

Specific health and safety concerns regarding NPs relate to their small size that allows the NPs to interact with biological entities in new ways, such as engaging with biological receptors thereby ensuring active uptake processes and access to sub-cellular organelles, or through binding of biological macromolecules such as proteins, lipids and polysaccharides. Such interactions are driven by the chemical composition of the NP (its' synthetic identity), and result in a context-dependent biological identity.(Liu *et al.*, 2011; Rallo *et al.*, 2011; Zhang *et al.*, 2012; Tantra *et al.*, 2015; Fourches *et al.*, 2016; Mu *et al.*, 2016; Fjodorova *et al.*, 2017) Additionally, some of the features of NPs themselves pose an inherent risk to living organisms, such as where there is the possibility for transfer of electrons from the NPs to the cells, in specific cases where the band-gap (the difference in energy between the valence band and the conduction band of a solid material that consists of the range of energy values forbidden to electrons in the material) of the NP overlaps with that of cells, which has been linked to a specific toxicity mechanism for metal oxide NPs.(Zhang *et al.*, 2012)

Nanoinformatics methods and techniques, developed to support safer NP design and risk assessment of NPs, have been rapidly advancing in recent years and are especially targeting the development of useful tools addressing the needs of regulators.(Bates *et al.*, 2015; David A. Winkler, 2016) Quite recently, a range of nanoinformatics tools have been proposed to assess the biological responses of different NPs that helped facilitate informed debate on how to best direct the ongoing efforts towards development of safe(r)-by-design (nano)materials. (Marvin *et al.*, 2013; Kleandrova *et al.*, 2014; Melagraki and Afantitis, 2014; Mikolajczyk *et al.*, 2015; Kar *et al.*, 2016; Tämm *et al.*, 2016; Isayev *et al.*, 2017).

For those NPs developed specifically for biomedical applications or consumer products and that have direct contact with humans, it is vital to understand the interaction of nanostructures with the biological environment. It has been experimentally demonstrated that when NPs enter a biological medium, the surface of NPs is selectively covered by different proteins forming the so called 'protein corona'.(Cedervall *et al.*, 2007; Lynch *et al.*, 2014; Vilanova *et al.*, 2016) The protein modified surface of

the NPs is then exposed to the surrounding environment affecting its subsequent interactions with biological entities (cells, organisms etc.). The interactions of NPs with proteins have been demonstrated to be sufficiently long-lived, with slow exchange times, such that the bare surface of a NP is never exposed.(Cedervall *et al.*, 2007; Monopoli *et al.*, 2012)

For each NP composition and form (shape, capping, charge etc.), this protein corona is uniquely formed, and its (ensemble) composition is dependent on the nature of the biological environment, and as such is context-dependent. The NP physicochemical parameters that crucially affect the structure and composition of the protein corona include size, shape, composition, hydrophobicity, surface modifiers and charges which influence both which proteins bind, and their bound conformation.(Hadjidemetriou and Kostarelos, 2017; García-Álvarez *et al.*, 2018) Further work is needed (experimental and theoretical) to understand the thresholds in terms of each physicochemical parameter leading to significant differences in corona composition – e.g. the effect of the width of the NP size distribution and mean size on corona composition. The ultimate goal would be to be able to predict the composition of the NP corona from its physico-chemical parameters, and ultimately then to predict its biological fate (uptake, localisation and impacts) on the basis of its corona, allowing classification and grouping of NPs.

The nature of the exposure environment and the duration of exposure also influence the final corona composition, and small changes in protein structure (single amino acid substitutions) have been shown to significantly alter corona thickness and stability.(Treuel *et al.*, 2014) Thus, the relative abundances of proteins identified in different NP protein coron*ae* do not directly correlate with their respective abundances in the biological medium. Note also that not all proteins in the corona are directly interacting with the NP surface; protein-protein interactions also play a vital role in determining the corona composition, adding to the complexity of understanding and predicting corona compositions.(Stigler *et al.*, 2010) Thus, utilisation of existing knowledge on protein-protein interactions, and incorporation of these interaction constants into modelling approaches would be a useful for prediction of NP protein coronas and their dynamics as NPs move into and potentially between cells. However, in order to develop predictive computational models, such as for NP proteins coronas and how the nature of the corona influences the cellular uptake and impact of the NPs, there is first a need for robust datasets against which to train and test the models.

Although several studies and reviews have been published to determine and describe the protein corona formation for a variety of NPs,(Pino *et al.*, 2014; Farrera and Fadeel, 2015; Foroozandeh and Aziz, 2015) including some limited efforts to correlate corona formation with biological efficiency in cells,(Treuel *et al.*, 2014; Varnamkhasti *et al.*, 2015), these data are generally disparate and insufficiently large to use for

model development and consequently efforts to computationally explore these data are very limited to date (Liu *et al.*, 2015; Bigdeli *et al.*, 2016). Recently, a library of gold NPs with different sizes and surface modification have been synthesized and tested in terms of their physicochemical properties, protein corona compositions and cell association, with the goal of developing a predictive model for NP cell association.(Walkey *et al.*, 2014) Data on protein corona were also explored and relative abundances for several proteins were determined for each NP. The inclusion of serum protein corona fingerprints to predict cell association using Partial Least Squares (PLS) regression, was proven more accurate than a model that used only physical information for the NPs. The authors thus concluded that protein corona encodes more biologically relevant information about a NP than its physical properties, and presented the idea of the corona as a "fingerprint" predictive of subsequent cellular behaviour. Moreover, the authors suggested that protein corona fingerprints can be extended to predict the association of NPs with other physiologically relevant cell types. In a newer publication the same authors also explored more linear as well as non-linear quantitative structure-activity relationships (QSARs) to derive important correlations for the prediction of cell association.(Liu *et al.*, 2015) However, work in this area is very new, and papers showing no correlation / predictive capacity from the corona for a biological effect can also be found – e.g. Dobrovolskaia et al. found that corona composition did not accurately predict hematocompatibility of colloidal gold NPs.(Dobrovolskaia *et al.*, 2014)

In this work, a nanoinformatics workflow was developed with the dual aim to propose a validated predictive model for NP cell association based on a set of significant descriptors and to facilitate the use of the model by allowing its free online access within a user friendly interface. The validated open access model to quantitatively define the cellular association of gold NPs based on their physicochemical properties combined with available data on their acquired protein corona from undiluted human serum utilised data from (Walkey *et al.*, 2014). The developed model was made publicly available through the Enalos Cloud Platform and thus can be easily accessed and used by experts and non-experts interested in the design of safe NPs and their applications in medicine and elsewhere. This web service is a significant step forward from our previously published work on web services related to NPs hazard and risk assessment, as data on protein corona were exploited in addition to the critical physicochemical properties that influence the biological effects of NPs. Using this new tool, researchers, industry and regulators will be able to assess the likely biological behaviour of functionalized Au NPs and design the optimal surface functionalization strategy and physicochemical parameters to enhance or minimize association of their NPs with cells. Given that corresponding input parameters (serum corona information and uptake data) are available, the extension to other types of NPs is also feasible and the reliability of the predictions will be provided via the domain of applicability of the model. Extension to other cell types is more challenging as it would demand, for a given set of NPs, data on several cell types

that would be computationally explored to afford meaningful correlations among the different routes of NP uptake presented by different cells.

2. Methods

2.1 QNAR development via KNIME workflow

During QNAR development the following steps are required: data preprocessing, variable selection, model development and validation and domain of applicability determination. All these steps were implemented within the KNIME (Konstanz Information Miner) platform which is a freely available and open source tool that is increasingly used for solving chemoinformatics problems (Berthold *et al.*, 2009). For this purpose, existing nodes were combined with our in-house Enalos KNIME nodes that execute several important operations including model validation performed by the Enalos Model Acceptability Criteria node and domain of applicability determination performed by the Enalos Domain – Similarity node and Enalos Domain – Leverage node.(Melagraki Afantitis, A., 2013; Varsou *et al.*, 2017) These nodes have been developed by NovaMechanics and are publicly available through the KNIME Community and the company's website (NovaMechanics Ltd, 2013) and have been described in detail in previous publications.(Melagraki and Afantitis, 2013; Melagraki *et al.*, 2017) A brief schematic description of our workflow is provided in Figure 1.

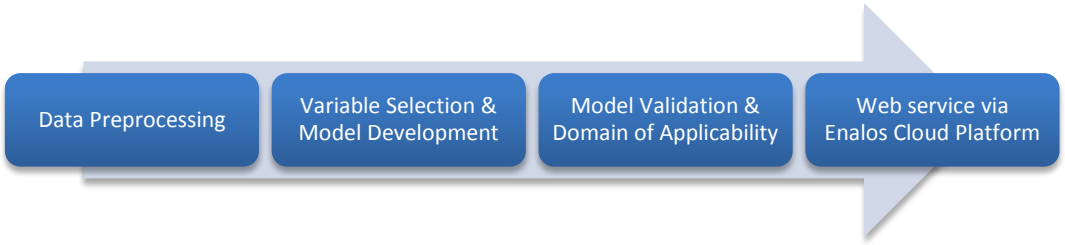


Figure 1. Nanoinformatics workflow for the development of NP cellular association predictive model.

Once the validated model was achieved, it was released as a ready-to-use application to be used in NP risk assessment by experts as well as non-experts. Our KNIME workflow enabled easy export of the model as a web service through the Enalos Cloud Platform as described below (Melagraki and Afantitis, 2014, 2015; NovaMechanics Ltd, 2017).

2.1.1 Data Set

We explored a data set that consists of 105 chemically diverse gold NPs with different surface modifiers.(Walkey *et al.*, 2014) Three different core sizes were included, namely 15, 30 and 60 nm. In

the formulations 67 organic surface modifiers were used, including small molecules, polymers, peptides, surfactants and lipids that can be characterized as “neutral”, “anionic” and “cationic” based on their chemical structure and net charge at physiological pH (pH 7.4). All NPs are coded based on their core size and surface modifier as shown in Walkey et. al.(Walkey *et al.*, 2014)

For each NP, several descriptors were available including physicochemical parameters measured after synthesis, and physicochemical parameters measured after exposure in undiluted human serum that stimulates the biomolecular environment with which the NP interacts during *in vitro* cell culture experiments. These parameters include: Surface area (cm^2/NP) based on TEM size, As synthesized and Serum Z-Average Hydrodynamic Diameters, Synthesized and Serum Volume Mean Hydrodynamic Diameters, As synthesized and Serum Number Mean Hydrodynamic Diameters, As synthesized and Serum Intensity Mean Hydrodynamic Diameters, As synthesized and Serum Polydispersity index, As synthesized and Serum Zeta Potential (mV), As synthesized and Serum-dispersed Localized Surface Plasmon Resonance (LSPR) index, As synthesized LSPR peak position (nm), Serum density Total protein (BCA assay), Total Au concentration (Au_{tot}, determined via ICP-AES) in nmol, Total surface area (SA_{tot}) in cm^2, Protein density (at the NP surface) in ug/cm^2.

On top of that, data on the presence of 785 distinct serum proteins across the entire library of NPs are also available. For each NP, distinct serum proteins were identified and quantified based on their relative protein spectral count and can be used as serum protein corona fingerprints for each formulation. On average, each NP formulation adsorbed 71 ± 22 distinct serum proteins. The abundances of the specific proteins attached to the different NPs were used as possible inputs for the QNAR model development. In total 805 parameters were extracted from the experimental dataset and included as possible inputs for the model development described here.

In parallel, data on NP association (which includes internalization of the NPs and adhesion to the cell membrane) with A549 human lung epithelial carcinoma cells in a monolayer culture, determined via inductively coupled plasma-atomic emission spectroscopy (ICP-AES), were also available. Experimental values of cell association for each NP included were calculated using the pseudopartition coefficient: $y = \frac{m_{cell}/m_{well}}{m_{cells}}$ where m_{cell} was the total atomic gold (or silver) content associated with cells, m_{well} was the total atomic gold (or silver) content in the exposure well (associated with cells and free in solution), and m_{cells} was the total mass of magnesium per sample. Thus, cell association data are expressed as the logarithm base of 2 expressed in units mL/ug (Mg) with values ranging from -11.123 to 1.327.(Walkey *et al.*, 2014; Liu *et al.*, 2015) The values of the variables considered as input parameters and the corresponding experimental values are given for each NP in Table S1 in the Supplementary file.

2.1.2 Model Development

KNIME provides the flexibility of exploring a great variety of methods for variable selection and modelling with minimum time required and thus several predictive KNIME workflows have been developed and applied in numerous cheminformatics / bioinformatics studies.(Nicola *et al.*, 2015; Steinmetz *et al.*, 2015; Yin *et al.*, 2015)

For our study, from the original pool of available variables, both physicochemical and protein fingerprints, a subset was first removed as some of the variables had low variance and added no discrimination power when included in the model. This was achieved by applying the 'Low Variance Filter' node straight after the data enter the KNIME workflow.

The Partitioning node included in KNIME was then used to divide the initial dataset into training and test sets in the ratio of 75:25 by applying the draw randomly option using the default random seed to get reproducible results upon re-execution of the node.

Among the available variables, a variable selection method was used to select the most important ones. Correlation–based feature subset selection (CfsSubset) variable selection, combined with BestFirst evaluator, were chosen to evaluate the most critical parameters for the training set.(Hall *et al.*, 2009; Witten, Frank and Hall, 2011). More details on these methods can be found in the Supporting Information.

In parallel with variable selection, a large variety of machine learning methods were explored to afford the variable combination that best suites the data. The approach considered here was the k Nearest Neighbour (kNN) methodology as the machine learning method combined with the variable selection method described above (CfsSubset) for performing regression to the available dataset. The kNN algorithm is a method for classifying objects based on closest training examples in the feature space and belongs to instance-based (or lazy) learning.(Zhang *et al.*, 2006) Based on the kNN algorithm, an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours (where k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbour. In this work, a distance weighted kNN algorithm was applied, Euclidean distance was used with all descriptors and contributions of neighbours weighted by the inverse of distance. The optimal k value (k=6 as per Table S2) was selected based on best model performance (Table S2).

2.1.3 Model Validation

The proposed model was fully validated both internally and externally as proposed by the Organization for Economic Cooperation and Development (OECD) principles for the validation, for regulatory purposes of QSARs (OECD, 2007) and represented by goodness-of-fit, robustness and predictivity as described below.(Zhang *et al.*, 2006; Puzyn *et al.*, 2017)

To evaluate the model performance, the following statistical criteria were used: the coefficient of determination between experimental values and model predictions (R^2), validation through an external test set, leave-many-out cross validation procedure and Quality of Fit and Predictive Ability of a continuous QSAR Model according to Tropsha's tests.(Melagraki *et al.*, 2007; Liu, Yao and Gramatica, 2009) Validation based on Tropsha's tests was made feasible by including the Enalos Model Acceptability Criteria node in our KNIME workflow. Details on the predictive ability formulas are given in the Supporting Information (equations S1-S5).

Moreover, a Y-randomization test was also used to ensure the robustness and the statistical significance of the predictive model. The dependent variable vector is randomly shuffled and a new model is developed using the original independent variable matrix. The derived models after several repetitions are expected to have less significant correlation coefficient values than the original model. This method is performed to eliminate the possibility of chance correlation. If the opposite happens then an acceptable predictive model cannot be obtained for the specific modelling method and data.

2.1.4 Domain of Applicability

When the model's limitations are known, predictions for any new entry can either be considered as reliable or unreliable based on some metrics compared to the model's limits. Structures (e.g. NP compositions) that fall outside the domain of applicability of the model are filtered out as the model cannot generate reliable predictions for these. The domain of applicability can be defined using similarity measurements based on the Euclidean distances among all training compounds and the test compounds or using leverages, as described in Supplementary Information. The Enalos Domain – Similarity node and the Enalos Domain – Leverage node that execute the aforementioned procedures are included in our workflow and were used to assess the domain of applicability of the proposed model.(Vrontaki *et al.*, 2015)

2.2 Enalos Cloud Platform

The Enalos Cloud Platform serves as a freely available web-based platform to support decision making by experts and non-experts regarding the biological activity, properties and toxicity of chemicals and NPs for

use in development of safe-by-design strategies and risk assessment. The Enalos Cloud Platform has been developed by NovaMechanics and is continuously extended as more validated models are developed.(Melagraki and Afantitis, 2014, 2015) The Enalos Cloud Platform has a user friendly format and was designed to address the needs of experts as well as non-experts.

Predictive workflows dedicated to the risk assessment of NPs have already been included in the Enalos Cloud Platform to act as a useful aid in the virtual screening of NPs. Models published in the Enalos Cloud platform are built based on reliable peer-reviewed and published data sources, and are based on the integration of advanced *in silico* tools to provide accurate predictions. The predictive model described in this work adds to our previous efforts on NPs safe-by-design and can be accessed at the following webpage: http://enalos.insilicotox.com/NanoProteinCorona/.

3. Results

3.1. Final Data Set for Model development

Application of the 'Low Variance Filter' node to the available data to remove variables that do not have any discrimination power (low variance),(OECD, 2007) resulted in all physicochemical parameters and 129 out of the 785 protein fingerprints remaining to be considered as input parameters for model development. The filtered dataset was pre-processed, normalized and randomly partitioned into training and validation set in a ratio of 75:25 using the Partitioning KNIME node by applying the default random seed. After data preprocessing, variable selection, model development, model validation and domain of applicability determination followed. Among the 105 NPs originally included in the dataset, 79 constituted the training set and 26 the validation or test set (see Table 1 where the validation set are indicated with *), of which 14 had 15nm cores, 5 had 30nm cores and 7 had 60nm cores, with 7 cationic, 5 neutral and 14 anionic surfaces. Only NPs included in the training set were used to develop the predictive model whereas NPs included in the test set were not involved by any means in the model development.

Table 1. NPs descriptions (size is indicated as G15, G30 or G60) including surface modification and charge (for full descriptions see the Supplementary information), Cellular Association (measured and predicted) and reliability of predictions for the validation test set (indicated with * in the ID column). From Walkey *et al.*, 2014; Liu *et al.*, 2015.

ID	NP		Cell Association (measured)	Cell Association (predicted)	Domain

			Log2 mL/μg (Mg)	Log2 mL/μg (Mg)	
1	G15.AC	Anionic	-5.183805957	-5,181354241	
2	G15.AHT	Cationic	-1.008545824	-1,105799078	
3	G15.Ala-SH	Anionic	-5.504706794	-5,539984404	
4	G15.Asn-SH	Anionic	-5.676379339	-5,667675173	
5	G15.AUT	Cationic	-1.315678187	-1,404784637	
6	G15.CALNN	Anionic	-7.137754348	-7,089063955	
7*	G15.CIT	Anionic	-5.419973023	-6,237281392	reliable
8*	G15.cPEG5K-SH	Neutral	-7.741605599	-7,529702202	reliable
9	G15.cPEG5K-SH (LD)	Neutral	-7.886871112	-7,806827755	
10*	G15.CTAB	Cationic	-5.862112795	-6,671601425	reliable
11	G15.DDT@BDHDA	Cationic	-7.293801253	-7,273753288	
12	G15.DDT@CTAB	Cationic	-7.589736344	-7,460190789	
13	G15.DDT@DOTAP	Cationic	-1.127551204	-1,183361387	
14	G15.DDT@ODA	Cationic	-6.121702418	-6,210757259	
15	G15.DDT@SA	Anionic	-6.803915548	-6,782461817	
16	G15.DDT@SDS	Anionic	-7.675057239	-7,574950302	
17	G15.DTNB	Anionic	-6.083455091	-5,992147102	
18	G15.F127	Anionic	-5.36093598	-5,471712175	
19	G15.Gly-SH	Anionic	-4.975181466	-5,005932259	
20	G15.HDA	Cationic	-0.270326496	-0,325215372	
21*	G15.LA	Anionic	-5.964253977	-6,565603538	reliable
22*	*G15.*MAA	Anionic	-6.142429499	-5,276864107	reliable
23	G15.MBA	Anionic	-5.381432273	-5,36369475	
24	G15.MES	Anionic	-3.199317327	-3,31486014	
25	G15.Met-SH	Anionic	-5.928482312	-5,912529871	
26	G15.MHA	Anionic	-5.73573998	-5,686345489	
27	G15.MHDA	Anionic	-5.778134503	-5,781290783	
28	G15.MPA	Anionic	-5.395652798	-5,396667548	
29	G15.mPEG1K-SH	Neutral	-10.75701655	-10,54636627	
30	G15.mPEG20K-SH (LD)	Neutral	-9.582595393	-9,526586909	
31	G15.mPEG2K-SH	Neutral	-7.857546161	-7,831881975	
32	G15.mPEG5K-SH	Neutral	-9.67146499	-9,601772313	
33	G15.mPEG5K(NH2)-SH	Neutral	-8.116498147	-8,099114047	
34	G15.MSA	Anionic	-6.10444776	-6,06976071	
35*	G15.MUA	Anionic	-4.846906601	-5,535995384	reliable
36*	G15.MUEG4	Neutral	-7.071905418	-6,708752158	unreliable
37	G15.MUTA	Cationic	0.11274716	0,097099997	
38	G15.nPEG5K-SH	Neutral	-10.99355273	-10,7320945	
39*	G15.nPEG5K-SH (LD)	Neutral	-6.068267192	-6,403093022	reliable
40	G15.NT@DCA	Anionic	-8.588936426	-8,529271912	
41	G15.NT@F127	Neutral	-6.426294194	-6,563387183	
42*	G15.NT@PSMA-AAP	Anionic	-6.223250188	-6,337957661	reliable
43	G15.NT@PSMA-EA	Anionic	-6.1682121	-6,198950134	
44	G15.NT@PSMA-EDA	Anionic	-5.403231085	-5,466962463	
45	G15.NT@PSMA-Urea	Anionic	-6.360621059	-6,389058819	
46	G15.NT@PVA	Neutral	-5.80079106	-5,86429014	
47	G15.ODA	Cationic	-2.765190582	-2,935815348	
48*	G15.PAH-SH	Cationic	-0.980042834	-1,285164209	reliable
49	G15.PEG3K(NH2)-SH	Neutral	-9.251353003	-9,188574234	
50	G15.PEI-SH	Cationic	-1.292169896	-1,320999634	
51*	G15.Phe	Anionic	-5.303711518	-6,482925693	reliable
52	G15.Phe-SH	Anionic	-6.268908041	-6,259990898	

53*	G15.PLL-SH	Cationic	-0.965765859	-0,84352245	reliable
54*	G15.PVA	Anionic	-7.929043536	-6,111843589	reliable
55	G15.PVP	Anionic	-6.022283964	-6,055637514	
56	G15.SA	Anionic	-5.916395245	-5,968381717	
57	G15.Ser-SH	Anionic	-5.224189072	-5,225632957	
58	G15.SPP	Anionic	-6.611425167	-6,554908695	
59	G15.T20	Anionic	-5.916658634	-5,941138028	
60	G15.Thr-SH	Anionic	-5.859606618	-5,870504099	
61	G15.TP	Anionic	-4.652498266	-4,676594598	
62*	G15.Trp-SH	Anionic	-6.965231933	-6,167321956	reliable
63	G30.AC	Anionic	-3.43591682	-3,526548823	
64	G30.AUT	Cationic	-1.630813246	-1,660831742	
65*	G30.CALNN	Anionic	-7.551188457	-5,429385858	reliable
66	G30.CFGAILS	Anionic	-6.356212428	-6,337558073	
67	G30.cPEG5K-SH	Neutral	-11.12261143	-11,02499026	
68	G30.DDT@BDHDA	Cationic	-4.95219456	-4,99220669	
69	G30.DDT@CTAB	Cationic	-7.598915524	-7,552880113	
70	G30.DDT@DOTAP	Cationic	-0.513598137	-0,505106609	
71*	G30.DDT@HDA	Cationic	-5.30868286	-5,676779587	reliable
72	G30.LA	Anionic	-5.051279148	-5,070347461	
73*	G30.MAA	Anionic	-6.142085992	-4,539660116	reliable
74	G30.Met-SH	Anionic	-5.22867711	-5,193784893	
75	G30.MHDA	Anionic	-4.32610319	-4,325047307	
76	G30.mPEG20K-SH (LD)	Neutral	-7.762460306	-7,752086318	
77	G30.MUA	Anionic	-4.555008456	-4,560375643	
78	G30.MUTA	Cationic	0.500053177	0,481741091	
79	G30.NT@F127	Neutral	-7.458863564	-7,464573452	
80	G30.PAH-SH	Cationic	-0.697714084	-0,713849845	
81*	G30.Thr-SH	Anionic	-5.737355739	-5,018852701	reliable
82*	G30.TP	Anionic	-4.142553328	-5,226708248	reliable
83	G60.AUT	Cationic	-0.779821601	-0,797637212	
84	G60.CIT	Anionic	-4.747717784	-4,790419222	
85	G60.cPEG5K-SH (LD)	Neutral	-6.23895505	-6,231556175	
86*	G60.CTAB	Cationic	-4.067031065	-4,480795555	reliable
87*	G60.CVVIT	Anionic	-4.647958051	-4,711362996	reliable
88*	G60.DDT@BDHDA	Cationic	-4.23739786	-3,91259778	reliable
89*	G60.DDT@DOTAP	Cationic	-0.303146847	-0,134886074	reliable
90	G60.DTNB	Anionic	-5.871960339	-5,786210906	
91	G60.HDA	Cationic	-1.008863835	-1,067392687	
92	G60.MBA	Anionic	-2.689194352	-2,763642933	
93	G60.MPA	Anionic	-3.07952409	-3,130724639	
94*	G60.mPEG20K-SH	Neutral	-8.786059054	-7,903638297	reliable
95	G60.mPEG5K-SH	Neutral	-10.22835958	-10,08667207	
96	G60.MUTA	Cationic	1.327288978	1,278470097	
97*	G60.nPEG5K-SH	Neutral	-7.181117171	-7,794940362	reliable
98	G60.NT@PSMA-AP	Anionic	-4.338543242	-4,380532584	
99	G60.NT@PVA	Neutral	-4.263386251	-4,377940842	
100	G60.ODA	Cationic	-3.370987013	-3,424311113	
101	G60.Phe-SH	Anionic	-4.14596268	-4,151055721	
102	G60.PVA	Anionic	-5.401348342	-5,33465243	
103	G60.Ser-SH	Anionic	-4.040817544	-4,058929332	
104*	G60.SPP	Anionic	-4.430879842	-4,427412398	reliable
105	G60.Trp-SH	Anionic	-3.954229393	-3,983932184	

*Test Set

Table 2. Data on the proteins found to predictive of gold NP uptake into A549 human lung epithelial carcinoma cells.

Protein UniCode	Protein description and role	Functional Annotation	Protein characteristics		Identified in proteins coronas of other NPs
			Molecular weight	Iseoelecctric point (IEP)	
P01024	**Complement C3** plays a central role in the activation of the complement system (both classical and alternative complement pathways). After activation C3b can bind covalently, via its reactive thioester, to cell surface carbohydrates or immune aggregates, and is considered to be opsonising protein enhancing NP uptake.(Scieszka *et al.*, 1991)	Complement system	188,688	5.96	PS NPs with various surface functionalisations (Ritz *et al.*, 2015) Ag NPs with citrate or PVP capping (20 & 110 nm citrate, 110 nm PVP) (Shannahan *et al.*, 2013)
P02766	**Transthyretin** is a thyroid hormone-binding protein which probably transports thyroxine from the bloodstream to the brain, as well as having a role in extracellular matrix organization. Also known as proalbumin.	Other Plasma components	16,001	5.40	SPIONS specifically dextran coated (Sakulkhu *et al.*, 2014) Doxorubicin-loaded dextran- stabilized PBCA NPs oated with poloxamer 188 or polysorbate 80 **in** rat plasma (Petri *et al.*, 2007) Solid lipid NPs (SLNs) coated with poloxamine 908 or poloxamer 407 (Göppert and Müller, 2005)
P08697	**Alpha-2-antiplasmin** is a serine protease inhibitor whose major targets are plasmin and trypsin, but it also inactivates chymotrypsin. It plays a role in blood coagulation and acute-phase response, as well as being involved on positive regulation of cell-cell adhesion	Coagulation	54,908	5.84	110nm AgNPs with citrate or PVP capping (Shannahan *et al.*, 2013)

	mediated by cadherin.				
P19823	**Inter-alpha-trypsin inhibitor heavy chain H2 m**ay act as a carrier of hyaluronan in serum or as a binding protein between hyaluronan and other matrix protein, including those on cell surfaces in tissues to regulate the localization, synthesis and degradation of hyaluronan which are essential to cells undergoing biological processes.	Other Plasma components	10,6920	6.40	110nm AgNPs with citrate or PVP capping (Shannahan *et al.*, 2013)
Q13103	**Secreted phosphoprotein 24** is involved in cellular protein metabolic process and has been identified in the extracellular exosome, suggesting a potential role in NP exocytosis.	Other Plasma components	24,623	8.39	polyvinyl-alcohol-coated SPIONs with various surface charges
Q9UK55	**Protein Z-dependent protease inhibitor** inhibits activity of the coagulation protease factor Xa in the presence of PROZ, calcium and phospholipids. Also inhibits factor XIa in the absence of cofactors.	Other Plasma components	50,821	8.55	Uncoated PLGA nanoparticles (Sempf *et al.*, 2013)
P02788	**Lactotransferrin** is an iron binding transport proteins which can bind two Fe^{3+} ions in association with the binding of an anion, usually bicarbonate. Binds specifically to the lipid A portion of bacterial lipopolysaccharide (LPS). Lipopolysaccharide-mediated signaling pathway and positive regulation of toll-like receptor 4 signaling pathway,	Tissue Leakage	80,064	8.01	Didn't find any papers other than the gold NPs on which the QSAR is based.

	suggestive of role in NP uptake.				
P02775	**Platelet basic protein** stimulates DNA synthesis, mitosis, glycolysis, intracellular cAMP accumulation, prostaglandin E2 secretion, and synthesis of hyaluronic acid and sulfated glycosaminoglycan.	Acute Phase	14,179	9.07	30, 200 and 400 nm Fe_3O_4 NPs (Hu *et al.*, 2014) In coronas of 6 of 17 tested Liposomes compositions / sizes (positive and negative) (Bigdeli *et al.*, 2016)
P14625	**Endoplasmin** is a molecular chaperone that functions in the processing and transport of secreted proteins via receptor-mediated endocytosis. Required for proper folding of Toll-like receptors (toll-like receptor signaling).	Tissue Leakage	92,754.15	4.56	Didn't find any papers other than the gold NPs on which the QSAR is based.
Q96KN2	**Beta-Ala-His dipeptidase** is involved in metal ion binding and regulation of cellular protein metabolic process including proteolysis.	Tissue Leakage	55,090	5.23	In coronas of 3 of 17 tested Liposomes compositions / sizes (positive and negative) (Bigdeli *et al.*, 2016)

3.2 Identification of Predictive Parameters

All available parameters (13 physicochemical parameters and 129 protein fingerprints) were evaluated for their ability to quantitatively describe the biological response, in this case association with A549 cells. All available data, including physicochemical data prior to and after exposure to blood serum and protein corona fingerprints, were combined to identify the subset of descriptors that best describes the variation of the measured NP-cell association. For this purpose, CfsSubset variable selection combined with BestFirst evaluator were applied on the training data to select the most significant descriptors. Among the available descriptors, 13 have emerged as the most critical in capturing the significant structural characteristics that affect the cellular association of the studied NPs, including the abundance of each of 10 proteins from the corona fingerprints (see Table 2 for details of the specific proteins and their known biological functions) and 3 physicochemical parameters, all of which are the parameters measured on the NPs dispersed in serum, i.e. the "with serum Z-Average Hydrodynamic Diameter", the "with serum Zeta Potential (mV)" and the "with serum (Au) Localized Surface Plasmon Resonance (LSPR) index". This is a really important insight, as it suggests that NP characterisation in the serum medium is much more predictive up uptake than the same parameters characterised in the absence of serum, again confirming the important role of the NP corona in driving NP association with cells and subsequent internalisation.

3.3 Validation of the predictive model

Given the selected parameters, the kNN methodology was selected to best correlate the input data with the observed biological response, i.e., cellular association. Given the flexibility of our KNIME workflow to test a great number of modelling methodologies with minimum time required, the proposed methods were identified as the combination that best describes our data and outperformed various different individual algorithms that were also tested.

To verify the model's robustness and accuracy several validation tests were performed as described in the Materials and Methods Section to address the principles recommended by the OECD including robust validation of results. Based on the results for the training and tests sets, predicted values for cellular association of each NP are included in Table 1. A summary of the produced results as extracted from the Enalos Model Validation node is given in Figure 2. As can be seen from the results, the significance, accuracy and robustness of the model are illustrated by the corresponding statistics shown in Figure 2.

Figure 2. Model Validation results

Criterion	Assessment	Result
R^2 > 0.6	PASS	R^2 = 0.832
Rcvext^2 > 0.5	PASS	Rcvext^2 = 0.83
(R^2 - R0^2)/R^2 < 0.1	PASS	(R^2 - R0^2)/R^2 = -0.199
(R^2 - R'0^2)/R^2 < 0.1	PASS	(R^2 - R'0^2)/R^2 = -0.179
abs(R0^2 - R'0^2) < 0.1	PASS	abs(R0^2 - R'0^2) = 0.016
0.85 < k < 1.15	PASS	k = 1.014
0.85 < k' < 1.15	PASS	k' = 0.965

Model Predictive

The "Leave ten out" (L10O) cross validation procedure was also applied to the available dataset and the model was proven to be quite stable to the inclusion-exclusion of data. The corresponding statistical value was measured equal to 67%. A Y-randomization test was additionally performed to further test the robustness and the statistical significance of the proposed model and eliminate the possibility of chance correlation. After applying this technique by randomly shuffling the response value multiple times no statistically significant models were retrieved.

3.4 Domain of Applicability

For all NPs included in the test set, the domain of applicability was defined based on both Euclidean distances and leverages, as described in the Methods section. This step describes the limitations of the model and undertakes the important work of highlighting the structures that cannot be tolerated by the model and thus indicates the predictions that can or cannot be considered reliable. The applicability domain limit value was measured equal to 3.338 based on equation S6 described in the supporting information. This value was compared to the calculated distance between a NP included in the test and its nearest neighbour in the training set. Among the NPs included in the test set, one (ID 36 in Table 1) had a value that exceeded the applicability domain limit and therefore the prediction for this NP cannot be considered reliable. All other NPs in the test set had values in the range of 0.326 - 2.98, less than the applicability domain limit, and therefore all predictions for these NPs fell inside the domain of applicability of the model and can thus be considered reliable.

In addition, domain of applicability was determined based on leverages based on equation (S7) described in supporting information. The leverages were plotted against the residuals for each NP and the Williams plot can also be seen in supporting information (Figure S1). Four NPs included in the training set and six NPs included in the test set were identified with leverages higher than the limit and are plotted on the right side of the leverage limit presented in the Figure. A consensus approach based on both methodologies can used to assess the reliability of model's predictions.

3.5 Provision of open access to the model via the Enalos Cloud Platform

Finally, the predictive NP cellular association model was made publicly available online through Enalos Cloud Platform (NovaMechanics Ltd, 2017). The Enalos Cloud platform is a web service based solely on open source and in house algorithms and software and was developed with the purpose of making QSAR models available to the interested user wishing to generate evidence on adverse effects in the decision making framework. We have developed a ready-to-use application based on our QNAR model using of this open source platform that already hosts other validated and predictive models that can be utilized in the NPs design process. In this way our model can be immediately explored by anyone interested in NP design. Enalos Cloud platform provides a user friendly interface with no special computational skills required and a procedure with minimum time required just for importing and submitting the input variables.

The web service is designed in a way that minimum steps are required to virtual screen a wide range of NPs. To initiate a prediction the user can either import the indicated parameters (i.e., 3 physicochemical descriptors measured in serum and 10 protein Spectral Counts as determined from the adsorbed corona from human serum) for a set of NPs or import a CSV file (.csv, see Figure 3 for a screenshot of the online platform) with several sets of properties included for High Throughput Virtual Screening. A prediction is then generated by clicking the submit button.

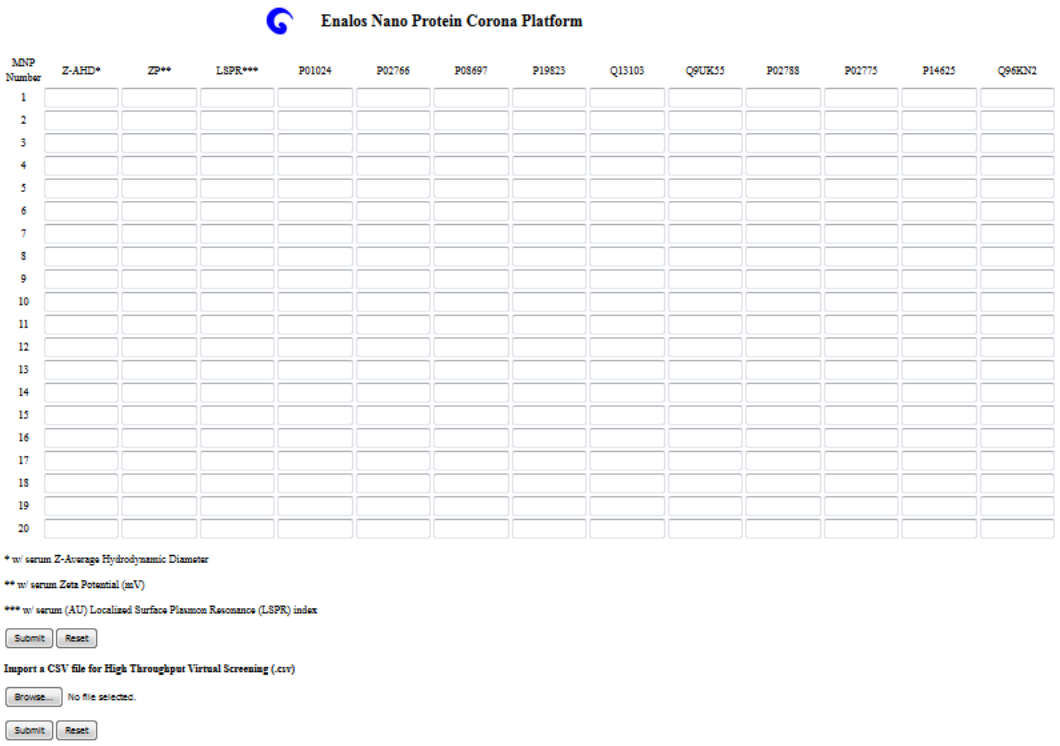
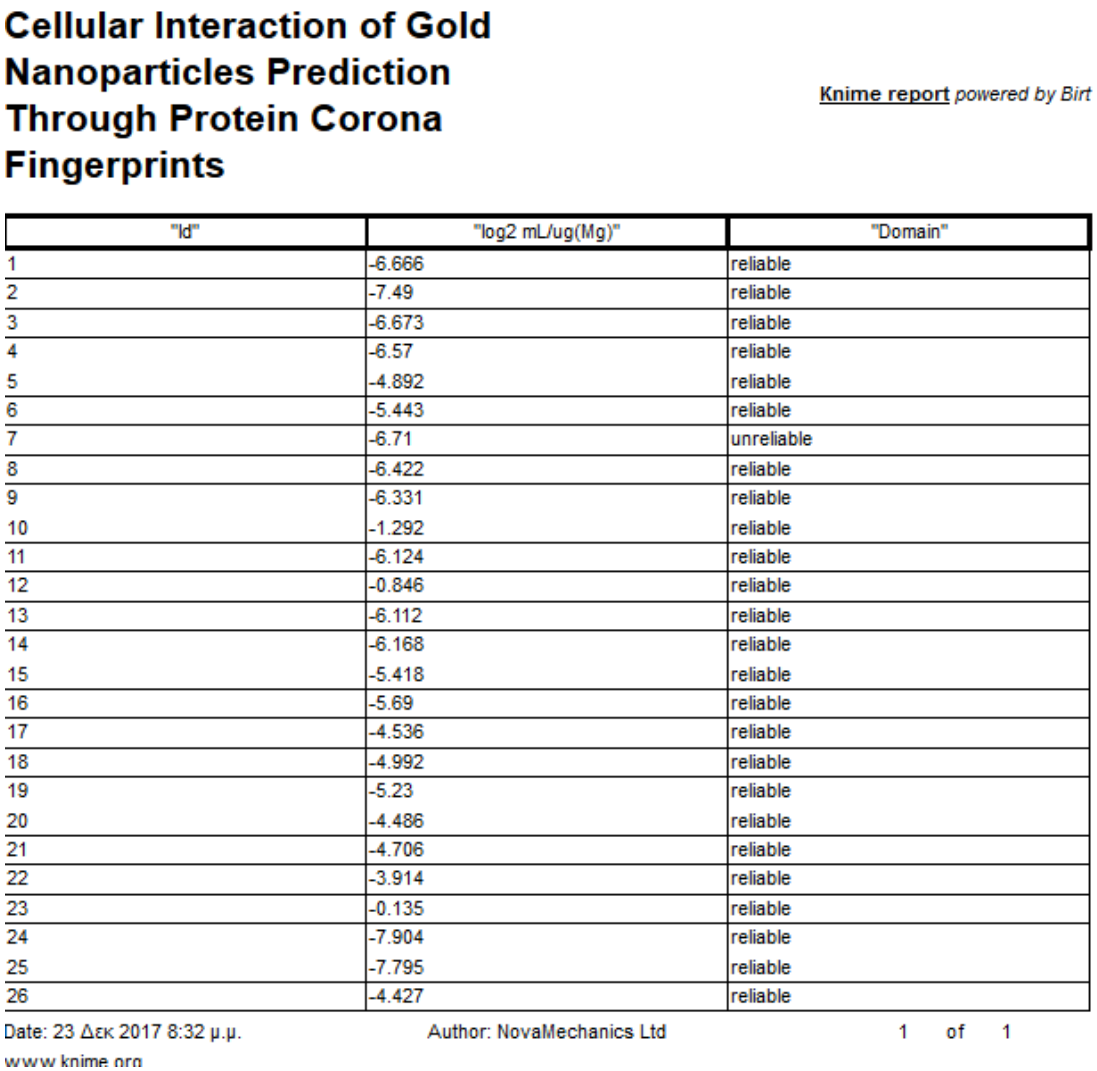


Figure 3. Screen Shot of Enalos Cloud Platform input page

The output is produced in the following formats: either as a summary of the results in a pdf like format on a different html page, or as a CSV file containing all the output information for further analysis. In each case, the results include the predicted value for each NP entered and an indication of whether the prediction can be considered reliable based on the domain of applicability of the model. A screen shot of the results page using the 26 NPs in the test set is presented in Figure 4. Note that the ID values in Figure 4 correspond to the test particles (with a *) from Table 1, numbered from 1-26 (i.e. G15.CIT to G60.SPP in order from Table 1). The Domain column tells if the prediction as reliable or not, and the log2 mL/µg(Mg) is the normalized cellular association of the NPs.



Cellular Interaction of Gold Nanoparticles Prediction Through Protein Corona Fingerprints

Knime report *powered by Birt*

"Id"	"log2 mL/ug(Mg)"	"Domain"
1	-6.666	reliable
2	-7.49	reliable
3	-6.673	reliable
4	-6.57	reliable
5	-4.892	reliable
6	-5.443	reliable
7	-6.71	unreliable
8	-6.422	reliable
9	-6.331	reliable
10	-1.292	reliable
11	-6.124	reliable
12	-0.846	reliable
13	-6.112	reliable
14	-6.168	reliable
15	-5.418	reliable
16	-5.69	reliable
17	-4.536	reliable
18	-4.992	reliable
19	-5.23	reliable
20	-4.486	reliable
21	-4.706	reliable
22	-3.914	reliable
23	-0.135	reliable
24	-7.904	reliable
25	-7.795	reliable
26	-4.427	reliable

Date: 23 Δεκ 2017 8:32 µ.µ. Author: NovaMechanics Ltd 1 of 1

www.knime.org

Figure 4. Screen Shot of Enalos Cloud Platform results. Note that log2 mL/µg(Mg) is the normalized cellular association of NPs, using the total magnesium (Mg) content to determine the total number of cells (see section 2.1.1 above for details). This method allows small changes in NP concentration (e.g. doubling) to be significant.

The web service can be easily used to produce a reliable prediction by implementing a one-step procedure that requires entering the requested properties, manually or as an CSV file, and submitting the task to initiate output generation. Any NP can be submitted individually or as part of a list of NPs submitted for High Throughput Virtual Screening. The predictions given by the developed model are generated within seconds after submission and the output appears as seen in Figure 4. The user can experiment with different values of the requested properties and study the characteristics that are responsible to induce a certain effect. The user can take advantage of the proposed QNAR model and immediately scan the NPs of interest for a preliminary *in silico* testing.

4. Discussion

Within this work we propose an *in silico* workflow that was developed in an effort to identify physicochemical properties and protein corona fingerprints that significantly correlate with NP cell association. Based on our findings, a predictive model was built that allows the prediction of cell association for a new given NP based on a reduced set of input (experimental) parameters. Our model was made publicly available through the Enalos Cloud Platform in order to maximize model's usability and facilitate virtual screening processes.

Experimental procedures for the biological evaluation of NPs are often costly and time-consuming, and full regulatory approval can often require 2-year animal studies. Indeed it has been estimated that the cost of preparing a REACH dossier for approval in the EU costs between €93 and €173 million per substance, prior to any modifications for NPs for which additional costs are required. These costs are dramatically increased in the absence of an extensive grouping and read-across approach for NPs reaching an additional cost between €100 and €600 million per NP (IHCP/2011/I/05/27/OC, 2013).

Thus, it is clear that alternative approaches based on computational methods should be proposed in the literature. Quantitative Nanostructure-Activity Relationships (QNARs) have recently emerged as a significant field of research for the prediction of the biological effects of NPs,(Rasulev *et al.*, 2012; Roca *et al.*, 2012; Kleandrova *et al.*, 2014; Speck-Planche *et al.*, 2015; David A Winkler, 2016; Vrontaki *et al.*, 2016; Toropova *et al.*, 2017) and several robust and predictive models have been proposed in the literature as highlighted in recent reviews.(Lynch *et al.*, 2017) However, lack of organized datasets, incoherent experimental data based on different protocols, and lack of available descriptors for nanostructures impose several restrictions that hamper progress and need to be addressed as a matter of priority. Well organized international efforts among regulatory agencies,

industry and academia [i.e. National Nanotechnology Initiative, NanoSafety Cluster] are currently ongoing have already been formed to work also towards this goal.

Validated QNAR models for NPs can significantly contribute to understanding of the underlying relationships(Gajewicz *et al.*, 2017) among structural characteristics and biological effects, but to achieve this, the produced models must be available in a user friendly format. If this is not taken into account, then the model cannot be directly explored by anyone interested. In this context, a key goal of this work has been building a robust and validated predictive model for the prediction of cell association of a large set of gold NPs based on their physicochemical characteristics and protein corona formed on their surface that is freely disseminated via a web service with a user friendly interface that can give online predictions for any given set of input parameters.

Our validated KNIME QNAR model, developed as described above, afforded a set of 13 variables that were selected as the most critical in describing the cell association of studied NMs, including 10 corona proteins and three physicochemical parameters. The corona proteins that were identified within the subset of critical descriptors are the following: P01024, P02766, P08697, P19823, Q13103, Q9UK55, P02788, P02775, P14625 and Q96KN2. Table 2 provides some information on the characteristics of these specific proteins (from UniProt), their known biological functions, and other NPs in whose coronas they have been identified. Many of these proteins have recognised cellular adhesion and/or transport functions carrying essential metals or other ligands in and out of cells. For example, Complement C3 (P01024) is a recognised opsonin (Scieszka *et al.*, 1991) (a molecule that binds to the surface of a foreign object marking it for phagocytosis by macrophages as part of the normal immune functioning), while Inter-alpha-trypsin inhibitor heavy chain (P19823) plays a role in binding to cell surfaces in tissues, and Alpha-2-antiplasmin (P08697) is involved in cell-cell adhesion. Thus, the presence of these proteins in the NP coronae correlating strongly with NP cellular adhesion is not surprising. Another group of the proteins identified in the NP coronae as correlating with NP cellular attachment have transport functions, meaning that the must be able to freely enter and exit cells. For example, Lesniak et al., demonstrated that NP uptake is a two-step process, where the NPs initially adhere to the cell membrane and are subsequently are internalized by the cells via energy-dependent pathways (Lesniak *et al.*, 2013). The authors also confirmed that the presence of a biomolecular corona confers specific interactions between the NP-corona complex and the cell surface including triggering of regulated cell uptake (Lesniak *et al.*, 2013). Thus, transthyretin (P02766), also known as proalbumin, transports hormones such as thyroxine to the brain, Beta-Ala-His dipeptidase (Q96KN2) is involved in metal ion transport, while Endoplasmin (P14625) is involved in the transport of secreted proteins via receptor-mediated endocytosis (as a molecular chaperone).

Secreted phosphoprotein 24 (Q13103) has been identified in the extracellular exosome suggesting a role in exocytosis, and Lactotransferrin (P02788) is an iron binding and transport protein as well as being an active component of the lipopolysaccharide-mediated signalling pathway and positive regulation of toll-like receptor 4 signalling pathway, suggestive of role in NP uptake. Neither Protein Z-dependent protease inhibitor (Q9UK55) not Platelet basic protein (P02775) have direct roles in cellular adhesion, being a protease inhibitor and stimulator of DNA synthesis, respectively, but as noted above not all proteins in the corona will be that outermost corona layer, nor all necessarily directly involved in the cellular attachment, since binding is based on affinity for the NP surface, and/or affinity for a protein already bound to the NP via protein-protein interactions.

Physicochemical parameters within the subset that are predictive were "with serum Z-Average Hydrodynamic Diameter", "with serum Zeta Potential (mV)" and the "with serum (Au) Localized Surface Plasmon Resonance (LSPR) index". These are quick and easy to measure, although their measurement in undiluted human serum (or the equivalent dilution of medium containing 10% foetal bovine serum proteins, so-called complete medium) is not always performed as standard in toxicological assays. Indeed, the cellular exposures in the dataset discussed here were conducted in RPMI medium supplemented with 10% foetal bovine serum while the corona data is for 100% human serum. Note also that LSPR measurements are only relevant for metal/metal oxide NPs, as this particular parameter is a consequence of quantum confinement of electrons in metal nanostructures. Whether the same predictivity would result from the use of corona data determined using undiluted bovine serum or the more typical 10% bovine serum utilized in cell culture remains to be evaluated. However, given the clear correlation between physicochemical characteristics in serum and cellular attachment demonstrated by the QNAR model, and the complete lack of correlation or predictive power from the equivalent characterisation in water, a clear recommendation from this work would be that experimentalists include measurements (size, zeta potential etc.) in the presence of the appropriate serum (i.e. human serum for human cell lines) as part of the routine physicochemical characterisation of their NPs, as they are most predictive of cellular association of the NPs. The value of this characterisation data in the appropriate exposure conditions is not just to facilitate modelling, but will also allow improved dose-response determinations, and increase the robustness of hazard and risk assessment based on experimental data, which if appropriately conducted can be used in weight of evidence arguments for regulatory registration dossiers, for example. While we agree that it's too easy to make recommendations to experimentalists, it is important to demonstrate both for enhanced modelling capacity but also for

accurate and meaningful risk assessment, the characterization of the NPs needs be performed under the conditions of the exposure in order to be able to make the sorts of clear dose-response correlations that are implicit in risk assessment but not always the actual case in NPs exposure.

The NP-cell association model presented here incorporates information available on NP biological interaction after exposure to the biological medium as given by different physicochemical parameters as well as protein corona data and was proven accurate and reliable for the given applicability limits. This model can be used to generate evidence on the biological response of NPs and could result in the reduction of costly and time consuming experiments for determining bioactivity, through focussing characterisation effort on the three physicochemical characteristics identified and a targeted search for the 10 proteins identified. Indeed, based on QNAR models of this type, there is scope for development of assays or antibody arrays to screen NP coronae for the presence of these specific proteins as a means to support data generation to feed into the QNAR models. While beyond the scope of the current article, this is an avenue we are interested to explore. Additionally, as more datasets appear, on other cell types or other NPs, we will continue to expand the domain of applicability of the model, and thus broaden its utility and predictive power.

Further, the web service can further add on screening existing databases or virtual chemical structures to identify NPs with desired properties. In this effort, the applicability domain will play an important role as it will filter out NPs that could not be tolerated by the model, for example, particles sizes > 70nm, or < 5nm, or differently shaped Au NMs such as rods, pyramids etc. Depending upon data availability, the model results could be further extended to different cell types and/or different serum sources or concentrations (bovine versus human or 10% versus undiluted).

4.1 User friendly interfaces through Enalos Cloud Platform

The dissemination of the developed predictive model to the wider community is a highly important aspect that is often neglected in the majority of examples presented in literature. In order to initiate further computational and experimental advances based on the produced results, it is of utmost importance that the proposed models do not remain within the developers' group but are immediately released in a friendly user format so that they can be easily explored in future NP design. On top of that, when open source and expandable tools are used, then the model's utility can be maximized since the workflow can be customized for the special needs of each project. Thus, our model was made available via Enalos Cloud Platform providing a user-friendly interface to enhance wider usability and acceptance.

The Enalos Cloud platform aspires to emerge as a useful tool to promote safer-by-design NPs and was created to boost developments in this direction by reducing the time and cost required for experimental evaluation. The user friendly interface makes this platform attractive for experts as well as non – experts and facilitates the *in silico* exploitation of available databases within a virtual screening framework to identify a prioritized list of NPs. (Melagraki Afantitis, A., 2013; Varsou *et al.*, 2017) Vrontaki *et al.*, 2016) In addition, the proposed workflows are easily expandable, and are adjustable to the specific needs of any other endpoint related to NPs adverse effects.

4. Conclusions

We have worked on one of the few extensive and well organized data sets correlating cellular attachment of gold NPs with their physicochemcial and acquired protein corona characteristics, and have presented the development of a fully validated and predictive QNAR model for cellular association. The QNAR is based on just 10 corona proteins and 3 physicochemical characteristics (determined in serum) that can be used in NPs design and virtual screening, reducing the amount of characterisation required for subsequent gold NPs that fall within the domain of applicability of the model. Our model was built based on open source and in house algorithms and models to perform all the crucial steps encountered, and is fully publicly available online through the Enalos Cloud platform to give immediate and easy access to the produced model and its results.

Cellular association was chosen as the end-point for the original experimental study due to its relevance to inflammatory responses, biodistribution, and toxicity *in vivo.* This model complements our previous efforts in developing *in silico* tools to promote the *in silico* exploration of the underlying correlations between nanostructures and biological effects and the development of safer-by-design NPs. It can also have application in the design of nanomedicines, via the identification of those NP physico-chemical parameters that will lead to uptake, cellular internalization and thus overcome one of the first key challenges of targeted delivery.

Acknowledgements

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 310451 (Project NanoMILE).

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Bates, M. E., Larkin, S., Keisler, J. M. and Linkov, I. (2015) 'How decision analysis can further nanoinformatics', *Beilstein Journal of Nanotechnology*, 6(1), pp. 1594–1600. doi: 10.3762/bjnano.6.162.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K. and Wiswedel, B. (2009) 'KNIME - The Konstanz Information Miner', *SIGKDD Explorations*, 11(1), pp. 26–31. doi: 10.1145/1656274.1656280.

Bigdeli, A., Palchetti, S., Pozzi, D., Hormozi-Nezhad, M. R., Baldelli Bombelli, F., Caracciolo, G. and Mahmoudi, M. (2016) 'Exploring Cellular Interactions of Liposomes Using Protein Corona Fingerprints and Physicochemical Properties', *ACS Nano*, 10(3), pp. 3723–3737. doi: 10.1021/acsnano.6b00261.

Cedervall, T., Lynch, I., Lindman, S., Berggård, T., Thulin, E., Nilsson, H., Dawson, K. A. and Linse, S. (2007) 'Understanding the nanoparticle-protein corona using methods to quantify exchange rates and affinities of proteins for nanoparticles.', *Proceedings of the National Academy of Sciences of the United States of America*, 104(7), pp. 2050–5. doi: 10.1073/pnas.0608582104.

D. Fourches, A. T. (2014) 'Quantitative Nanostructure-Activity Relationships: from unstructured data to predictive models for designing nanomaterials with controlled properties', in Monteiro-Riviere Tran, C.L., N. A. (ed.) *Nanotoxicology: Progress toward Nanomedicine*. CRC Press.

Dobrovolskaia, M. A., Neun, B. W., Man, S., Ye, X., Hansen, M., Patri, A. K., Crist, R. M. and McNeil, S. E. (2014) 'Protein corona composition does not accurately predict hematocompatibility of colloidal gold nanoparticles.', *Nanomedicine : nanotechnology, biology, and medicine*, 10(7), pp. 1453–63. doi: 10.1016/j.nano.2014.01.009.

Farrera, C. and Fadeel, B. (2015) 'It takes two to tango: Understanding the interactions between engineered nanomaterials and the immune system.', *European journal of pharmaceutics and biopharmaceutics : official journal of Arbeitsgemeinschaft fur Pharmazeutische Verfahrenstechnik e.V*, 95(Pt A), pp. 3–12. doi: 10.1016/j.ejpb.2015.03.007.

Fjodorova, N., Novic, M., Gajewicz, A. and Rasulev, B. (2017) 'The way to cover prediction for cytotoxicity for all existing nano-sized metal oxides by using neural network method', *Nanotoxicology*, 11(4), pp. 475–483. doi: 10.1080/17435390.2017.1310949.

Foroozandeh, P. and Aziz, A. A. (2015) 'Merging Worlds of Nanomaterials and Biological Environment: Factors Governing Protein Corona Formation on Nanoparticles and Its Biological Consequences', *Nanoscale Research Letters*. New York: Springer US, 10, p. 221. doi: 10.1186/s11671-015-0922-3.

Fourches, D., Pu, D., Li, L., Zhou, H., Mu, Q., Su, G., Yan, B. and Tropsha, A. (2016) 'Computer-aided design of carbon nanotubes with the desired bioactivity and safety profiles', *Nanotoxicology*, 10(3), pp. 374–383. doi: 10.3109/17435390.2015.1073397.

Gajewicz, A., Puzyn, T., Odziomek, K., Urbaszek, P., Haase, A., Riebeling, C., Luch, A., Irfan, M. A., Landsiedel, R., van der Zande, M. and Bouwmeester, H. (2017) 'Decision tree models to classify nanomaterials according to the *DF4nanoGrouping* scheme', *Nanotoxicology*. Taylor & Francis, pp. 1–17. doi: 10.1080/17435390.2017.1415388.

García-Álvarez, R., Hadjidemetriou, M., Sánchez-Iglesias, A., Liz-Marzán, L. M. and Kostarelos, K. (2018) 'In vivo formation of protein corona on gold nanoparticles. The effect of their size and shape.', *Nanoscale*, 10(3), pp. 1256–1264. doi: 10.1039/c7nr08322j.

Göppert, T. M. and Müller, R. H. (2005) 'Adsorption kinetics of plasma proteins on solid lipid nanoparticles for drug targeting.', *International journal of pharmaceutics*, 302(1–2), pp. 172–86. doi: 10.1016/j.ijpharm.2005.06.025.

Haase, A., Tentschert, J. and Luch, A. (2012) 'Nanomaterials: A Challenge for Toxicological Risk Assessment?', in *EXS*, pp. 219–250. doi: 10.1007/978-3-7643-8340-4_8.

Hadjidemetriou, M. and Kostarelos, K. (2017) 'Nanomedicine: Evolution of the nanoparticle corona.', *Nature nanotechnology*, 12(4), pp. 288–290. doi: 10.1038/nnano.2017.61.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009) 'The WEKA Data Mining Software: An Update', *ACM SIGKDD Explorations*, 11(1), pp. 10–18. doi: 10.1145/1656274.1656278.

Hu, Z., Zhang, H., Zhang, Y., Wu, R. and Zou, H. (2014) 'Nanoparticle size matters in the formation of plasma protein coronas on Fe3O4 nanoparticles.', *Colloids and surfaces. B, Biointerfaces*, 121, pp. 354–61. doi: 10.1016/j.colsurfb.2014.06.016.

IHCP/2011/I/05/27/OC (2013) *Examination and assessment of consequences for industry, consumers, human health and the environment of possible options for changing the REACH requirements for nanomaterials*. Edited by http://ec.europa.eu/environment/chemicals/nanotech/pdf/Final_Report.pdf.

Isayev, O., Oses, C., Toher, C., Gossett, E., Curtarolo, S. and Tropsha, A. (2017) 'Universal fragment descriptors for predicting properties of inorganic crystals.', *Nature communications*, 8, p. 15679. doi: 10.1038/ncomms15679.

Kar, S., Gajewicz, A., Roy, K., Leszczynski, J. and Puzyn, T. (2016) 'Extrapolating between toxicity endpoints of metal oxide nanoparticles: Predicting toxicity to Escherichia coli and human keratinocyte cell line (HaCaT) with Nano-QTTR', *Ecotoxicology and Environmental Safety*, 126, pp. 238–244. doi: 10.1016/j.ecoenv.2015.12.033.

Kleandrova, V. V., Luan, F., González-Díaz, H., Ruso, J. M., Speck-Planche, A. and Cordeiro, M. N. D. S. (2014) 'Computational tool for risk assessment of nanomaterials: novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions.', *Environmental science & technology*, 48(24), pp. 14686–94. doi:

10.1021/es503861x.
Lesniak, A., Salvati, A., Santos-Martinez, M. J., Radomski, M. W., Dawson, K. A. and ??berg, C. (2013) 'Nanoparticle adhesion to the cell membrane and its effect on nanoparticle uptake efficiency', *Journal of the American Chemical Society*, 135(4), pp. 1438–1444. doi: 10.1021/ja309812z.
Liu, H., Yao, X. and Gramatica, P. (2009) 'The Applications of Machine Learning Algorithms in the Modeling of Estrogen-Like Chemicals', *Combinatorial Chemistry & High Throughput Screening*, 12(5), pp. 490–496. doi: 10.2174/138620709788489037.
Liu, R., Jiang, W., Walkey, C. D., Chan, W. C. W. and Cohen, Y. (2015) 'Prediction of nanoparticles-cell association based on corona proteins and physicochemical properties', *Nanoscale*, 7(21), pp. 9664–9675. doi: 10.1039/C5NR01537E.
Liu, R., Rallo, R., George, S., Ji, Z., Nair, S., Nel, A. E. and Cohen, Y. (2011) 'Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles.', *Small (Weinheim an der Bergstrasse, Germany)*, 7(8), pp. 1118–26. doi: 10.1002/smll.201002366.
Lynch, I., Afantitis, A., Leonis, G., Melagraki, G. and Valsami-Jones, E. (2017) 'Strategy for Identification of Nanomaterials' Critical Properties Linked to Biological Impacts: Interlinking of Experimental and Computational Approaches', in *Advances in QSAR Modeling*. Springer, Cham, pp. 385–424. doi: 10.1007/978-3-319-56850-8_10.
Lynch, I., Dawson, K. A., Lead, J. R. and Valsami-Jones, E. (2014) 'Macromolecular Coronas and Their Importance in Nanotoxicology and Nanoecotoxicology', in *Frontiers of Nanoscience*, pp. 127–156. doi: 10.1016/B978-0-08-099408-6.00004-9.
Marvin, H. J. P., Bouwmeester, H., Bakker, M., Kroese, E. D., van de Meent, D., Bourgeois, F., Lokers, R., van der Ham, H. and Verhelst, L. (2013) 'Exploring the development of a decision support system (DSS) to prioritize engineered nanoparticles for risk assessment', *Journal of Nanoparticle Research*, 15(8), p. 1839. doi: 10.1007/s11051-013-1839-3.
Melagraki, G. and Afantitis, A. (2013) 'Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium', *Chemometrics and Intelligent Laboratory Systems*, 123. doi: 10.1016/j.chemolab.2013.02.003.
Melagraki, G. and Afantitis, A. (2014) 'Enalos InSilicoNano platform: an online decision support tool for the design and virtual screening of nanoparticles', *RSC Adv.*, 4(92), pp. 50713–50725. doi: 10.1039/C4RA07756C.
Melagraki, G. and Afantitis, A. (2015) 'A risk assessment tool for the virtual screening of metal oxide nanoparticles through enalos insiliconano platform', *Current Topics in Medicinal Chemistry*, 15(18), pp. 1827–1836. doi: 10.2174/1568026615666150506144536.
Melagraki, G., Afantitis, A., Sarimveis, H., Koutentis, P. a., Markopoulos, J. and Igglessi-Markopoulou, O. (2007) 'A novel QSPR model for predicting θ (lower critical solution temperature) in polymer

solutions using molecular descriptors.', *Journal of molecular modeling*, 13(1), pp. 55–64. doi: 10.1007/s00894-006-0125-z.

Melagraki, G., Ntougkos, E., Rinotas, V., Papaneophytou, C., Leonis, G., Mavromoustakos, T., Kontopidis, G., Douni, E., Afantitis, A. and Kollias, G. (2017) 'Cheminformatics-aided discovery of small-molecule Protein-Protein Interaction (PPI) dual inhibitors of Tumor Necrosis Factor (TNF) and Receptor Activator of NF-κB Ligand (RANKL).', *PLoS computational biology*, 13(4), p. e1005372. doi: 10.1371/journal.pcbi.1005372.

Melagraki Afantitis, A., G. (2013) 'Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium', *Chemometrics and Intelligent Laboratory Systems*, 123, pp. 9–14.

Mikolajczyk, A., Gajewicz, A., Rasulev, B., Schaeublin, N., Maurer-Gardner, E., Hussain, S., Leszczynski, J. and Puzyn, T. (2015) 'Zeta Potential for Metal Oxide Nanoparticles: A Predictive Model Developed by a Nano-Quantitative Structure–Property Relationship Approach', *Chemistry of Materials*, 27(7), pp. 2400–2407. doi: 10.1021/cm504406a.

Monopoli, M. P., Aberg, C., Salvati, A. and Dawson, K. A. (2012) 'Biomolecular coronas provide the biological identity of nanosized materials.', *Nature nanotechnology*, 7(12), pp. 779–86. doi: 10.1038/nnano.2012.207.

Mu, Y., Wu, F., Zhao, Q., Ji, R., Qie, Y., Zhou, Y., Hu, Y., Pang, C., Hristozov, D., Giesy, J. P. and Xing, B. (2016) 'Predicting toxic potencies of metal oxide nanoparticles by means of nano-QSARs', *Nanotoxicology*, 10(9), pp. 1207–1214. doi: 10.1080/17435390.2016.1202352.

Nicola, G., Berthold, M. R., Hedrick, M. P. and Gilson, M. K. (2015) 'Connecting proteins with drug-like compounds: Open source drug discovery workflows with BindingDB and KNIME.', *Database : the journal of biological databases and curation*. Oxford University Press, 2015. doi: 10.1093/database/bav087.

NovaMechanics Ltd (2013) *Enalos KNIME Nodes*. Available at: http://enalosplus.novamechanics.com (Accessed: 29 December 2017).

NovaMechanics Ltd (2017) *Enalos Cloud Platform*. Available at: http://www.insilicotox.com/index.php/products/predictive-models-web-services (Accessed: 29 December 2017).

OECD (2007) *Principles for the validation, for regulatory purposes of (Quantitative) Structure Activity Relationship Models*. Available at: http://www.oecd.org/env/ehs/oecdquantitativestructure-activityrelationshipsprojectqsars.htm.

Petri, B., Bootz, A., Khalansky, A., Hekmatara, T., Müller, R., Uhl, R., Kreuter, J. and Gelperina, S. (2007) 'Chemotherapy of brain tumour using doxorubicin bound to surfactant-coated poly(butyl cyanoacrylate) nanoparticles: revisiting the role of surfactants.', *Journal of controlled release : official journal of the Controlled Release Society*, 117(1), pp. 51–8. doi: 10.1016/j.jconrel.2006.10.015.

Pino, P. del, Pelaz, B., Zhang, Q., Maffre, P., Nienhaus, G. U. and Parak, W. J. (2014) 'Protein corona formation around nanoparticles – from the past to the future', *Mater. Horiz.*, 1(3), pp. 301–313. doi: 10.1039/C3MH00106G.

Puzyn, T., Jeliazkova, N., Sarimveis, H., Marchese Robinson, R. L., Lobaskin, V., Rallo, R., Richarz, A. N., Gajewicz, A., Papadopulos, M. G., Hastings, J., Cronin, M. T. D., Benfenati, E. and Fernández, A. (2017) 'Perspectives from the NanoSafety Modelling Cluster on the validation criteria for (Q)SAR models used in nanotechnology', *Food and Chemical Toxicology*. doi: 10.1016/j.fct.2017.09.037.

Rallo, R., France, B., Liu, R., Nair, S., George, S., Damoiseaux, R., Giralt, F., Nel, A., Bradley, K. and Cohen, Y. (2011) 'Self-organizing map analysis of toxicity-related cell signaling pathways for metal and metal oxide nanoparticles.', *Environmental science & technology*, 45(4), pp. 1695–702. doi: 10.1021/es103606x.

Rasulev, B., Gajewicz, A., Puzyn, T., Leszczynska, D. and Leszczynski*, J. (2012) 'Chapter 10. Nano-QSAR: Advances and Challenges', in, pp. 220–256. doi: 10.1039/9781849735476-00220.

Ritz, S., Schöttler, S., Kotman, N., Baier, G., Musyanovych, A., Kuharev, J., Landfester, K., Schild, H., Jahn, O., Tenzer, S. and Mailänder, V. (2015) 'Protein Corona of Nanoparticles: Distinct Proteins Regulate the Cellular Uptake', *Biomacromolecules*, 16(4), pp. 1311–1321. doi: 10.1021/acs.biomac.5b00108.

Roca, C. P., Rallo, R., Fernandez, A. and Giralt, F. (2012) 'Chapter 6 Nanoinformatics for Safe-by-Design Engineered Nanomaterials', *Towards Efficient Designing of Safe Nanomaterials: Innovative Merge of Computational Approaches and Experimental Techniques*, pp. 89–107. doi: 10.1039/9781849735476-00089.

Sakulkhu, U., Mahmoudi, M., Maurizi, L., Salaklang, J. and Hofmann, H. (2014) 'Protein corona composition of superparamagnetic iron oxide nanoparticles with various physico-Chemical properties and coatings', *Scientific Reports*, 4. doi: 10.1038/srep05020.

Scieszka, J. F., Maggiora, L. L., Wright, S. D. and Cho, M. J. (1991) 'Role of Complements C3 and C5 in the Phagocytosis of Liposomes by Human Neutrophils', *Pharmaceutical Research: An Official Journal of the American Association of Pharmaceutical Scientists*, 8(1), pp. 65–69. doi: 10.1023/A:1015830306839.

Sempf, K., Arrey, T., Gelperina, S., Schorge, T., Meyer, B., Karas, M. and Kreuter, J. (2013) 'Adsorption of plasma proteins on uncoated PLGA nanoparticles.', *European journal of pharmaceutics and biopharmaceutics : official journal of Arbeitsgemeinschaft fur Pharmazeutische Verfahrenstechnik e.V*, 85(1), pp. 53–60. doi: 10.1016/j.ejpb.2012.11.030.

Shannahan, J. H., Lai, X., Ke, P. C., Podila, R., Brown, J. M. and Witzmann, F. A. (2013) 'Silver Nanoparticle Protein Corona Composition in Cell Culture Media', *PLoS ONE*, 8(9). doi: 10.1371/journal.pone.0074001.

Speck-Planche, A., Kleandrova, V. V, Luan, F. and Cordeiro, M. N. D. S. (2015) 'Computational modeling in nanomedicine: prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model.', *Nanomedicine (London, England)*, 10(2), pp. 193–204. doi: 10.2217/nnm.14.96.

Steinmetz, F. P., Mellor, C. L., Meinl, T. and Cronin, M. T. D. (2015) 'Screening Chemicals for Receptor-Mediated Toxicological and Pharmacological Endpoints: Using Public Data to Build Screening Tools within a KNIME Workflow', *Molecular Informatics*. Wiley-VCH Verlag, 34(2–3), pp. 171–178. doi: 10.1002/minf.201400188.

Stigler, J., Lundqvist, M., Cedervall, T., Dawson, K. and Lynch, I. (2010) 'Protein Interactions with Microballoons: Consequences for Biocompatibility and Application as Contrast Agents', in *Ultrasound Contrast Agents*. Milano: Springer Milan, pp. 53–66. doi: 10.1007/978-88-470-1494-7_5.

Tämm, K., Sikk, L., Burk, J., Rallo, R., Pokhrel, S., Mädler, L., Scott-Fordsmand, J. J., Burk, P. and Tamm, T. (2016) 'Parametrization of nanoparticles: development of full-particle nanodescriptors', *Nanoscale*, 8(36), pp. 16243–16250. doi: 10.1039/C6NR04376C.

Tantra, R., Oksel, C., Puzyn, T., Wang, J., Robinson, K. N., Wang, X. Z., Ma, C. Y. and Wilkins, T. (2015) 'Nano(Q)SAR: Challenges, pitfalls and perspectives', *Nanotoxicology*, 9(5), pp. 636–642. doi: 10.3109/17435390.2014.952698.

Toropova, A. P., Toropov, A. A., Leszczynska, D. and Leszczynski, J. (2017) 'CORAL and Nano-QFAR: Quantitative feature – Activity relationships (QFAR) for bioavailability of nanoparticles (ZnO, CuO, Co3O4, and TiO2)', *Ecotoxicology and Environmental Safety*, 139, pp. 404–407. doi: 10.1016/j.ecoenv.2017.01.054.

Treuel, L., Brandholt, S., Maffre, P., Wiegele, S., Shang, L. and Nienhaus, G. U. (2014) 'Impact of protein modification on the protein corona on nanoparticles and nanoparticle-cell interactions', *ACS Nano*, 8(1), pp. 503–513. doi: 10.1021/nn405019v.

Tropsha, A., Mills, K. C. and Hickey, A. J. (2017) 'Reproducibility, sharing and progress in nanomaterial databases', *Nature Nanotechnology*, 12(12), pp. 1111–1114. doi: 10.1038/nnano.2017.233.

Varnamkhasti, B. S., Hosseinzadeh, H., Azhdarzadeh, M., Vafaei, S. Y., Esfandyari-Manesh, M., Mirzaie, Z. H., Amini, M., Ostad, S. N., Atyabi, F. and Dinarvand, R. (2015) 'Protein corona hampers targeting potential of MUC1 aptamer functionalized SN-38 core-shell nanoparticles.', *International journal of pharmaceutics*, 494(1), pp. 430–44. doi: 10.1016/j.ijpharm.2015.08.060.

Varsou, D. D., Melagraki, G., Sarimveis, H. and Afantitis, A. (2017) 'MouseTox: An online toxicity assessment tool for small molecules through Enalos Cloud platform', *Food and Chemical Toxicology*, 110, pp. 83–93. doi: 10.1016/j.fct.2017.09.058.

Vilanova, O., Mittag, J. J., Kelly, P. M., Milani, S., Dawson, K. A., Rädler, J. O. and Franzese, G. (2016) 'Understanding the Kinetics of Protein-Nanoparticle Corona Formation', *ACS Nano*, 10(12), pp.

10842–10850. doi: 10.1021/acsnano.6b04858.

Vrontaki, E., Mavromoustakos, T., Melagraki, G. and Afantitis, A. (2016) 'Quantitative Nanostructure-Activity Relationship Models for the Risk Assessment of NanoMaterials', in *Pharmaceutical Sciences*. IGI Global, pp. 1314–1338. doi: 10.4018/978-1-5225-1762-7.ch050.

Vrontaki, E., Melagraki, G., Mavromoustakos, T. and Afantitis, A. (2015) 'Exploiting ChEMBL database to identify indole analogs as HCV replication inhibitors.', *Methods (San Diego, Calif.)*, 71(C), pp. 4–13. doi: 10.1016/j.ymeth.2014.03.021.

Walkey, C. D., Olsen, J. B., Song, F., Liu, R., Guo, H., Olsen, D. W. H., Cohen, Y., Emili, A. and Chan, W. C. W. (2014) 'Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles.', *ACS nano*. American Chemical Society, 8(3), pp. 2439–55. doi: 10.1021/nn406018q.

Winkler, D. A. (2016) 'Recent advances, and unresolved issues, in the application of computational modelling to the prediction of the biological effects of nanomaterials.', *Toxicology and applied pharmacology*, 299, pp. 96–100. doi: 10.1016/j.taap.2015.12.016.

Winkler, D. A. (2016) 'Recent advances, and unresolved issues, in the application of computational modelling to the prediction of the biological effects of nanomaterials.', *Toxicology and applied pharmacology*, 299, pp. 96–100. doi: 10.1016/j.taap.2015.12.016.

Witten, I. H., Frank, E. and Hall, M. a. (2011) *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, *Annals of Physics*. doi: 10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Yin, Y., Xu, C., Gu, S., Li, W., Liu, G. and Tang, Y. (2015) 'Quantitative Regression Models for the Prediction of Chemical Properties by an Efficient Workflow', *Molecular Informatics*, 34(10), pp. 679–688. doi: 10.1002/minf.201400119.

Zhang, H., Ji, Z., Xia, T., Meng, H., Low-Kam, C., Liu, R., Pokhrel, S., Lin, S., Wang, X., Liao, Y.-P., Wang, M., Li, L., Rallo, R., Damoiseaux, R., Telesca, D., Mädler, L., Cohen, Y., Zink, J. I. and Nel, A. E. (2012) 'Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation.', *ACS nano*, 6(5), pp. 4349–68. doi: 10.1021/nn3010087.

Zhang, S., Golbraikh, A., Oloff, S., Kohn, H. and Tropsha, A. (2006) 'A novel Automated Lazy Learning QSAR (ALL-QSAR) approach: Method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models', *Journal of Chemical Information and Modeling*, 46(5), pp. 1984–1995. doi: 10.1021/ci060132x.